PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-134364
(43)Date of publication of application : 21.05.1999

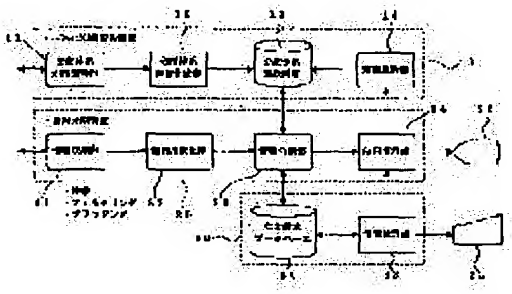
(51)Int.Cl. G06F 17/30

(21)Application number : 09-316245
(22)Date of filing : 31.10.1997

(71)Applicant : OMRON CORP
(72)Inventor : GO ATOU
FUJII FUJIKI
SAKAGUCHI MANABU
SOGO TAIJI
SAWADA AKIRA

(54) SYSTEMATIZED KNOWLEDGE ANALYZING METHOD AND DEVICE THEREFOR, AND CLASSIFYING METHOD AND DEVICE THEREFOR

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a systematized knowledge analyzing device and classifying device for extracting a prescribed terminal class suited to the content of an unclassified document, and relating them even when the state of structured data is not sufficiently known.
SOLUTION: In a systematized knowledge analyzing device 10, existing structured data and document information related with this are obtained, while a keyword extraction processing is operated to a document belonging to the same terminal class of the structured data, and a feature vector constituted of a significant word and weight is generated by a knowledge system dictionary preparing part 12, and the feature vector is stored as the feature of the terminal class with the obtained information in a classifying system knowledge dictionary 13. At the time of obtaining an unclassified document, the keyword extraction processing is operated, and the feature vector is generated by an information abstracting part 22 of an automatic classifying device 20, and the matching of the feature vector with the preliminarily registered feature vector of each terminal class is operated by an information classifying part 23, and allocation to the terminal class whose matching level is high is operated.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－134364

(43)公開日　平成11年(1999)５月21日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/40	３４０
			３７０Ａ
		15/401	３１０Ｄ

審査請求　未請求　請求項の数９　ＦＤ　（全 11 頁）

(21)出願番号　特願平9－316245

(22)出願日　平成９年(1997)10月31日

(71)出願人　000002945
オムロン株式会社
京都府京都市右京区花園土堂町10番地

(72)発明者　呉　亜棟
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(72)発明者　藤居　藤樹
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(72)発明者　坂口　学
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(74)代理人　弁理士　松井　伸一

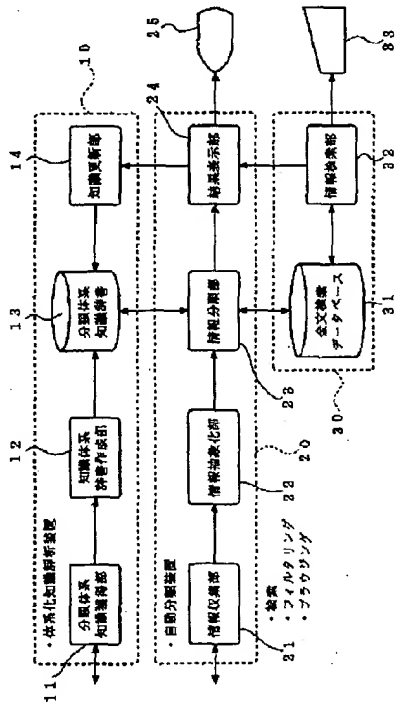
最終頁に続く

(54)【発明の名称】　体系化知識解析方法及び装置並びに分類方法及び装置

(57)【要約】

【課題】　構造化データの状態について十分に知らなくても、未分類の書類をその内容にあった所定の末端クラスを抽出し、関連づけることができる体系化知識解析装置及び分類装置を提供すること

【解決手段】　体型化知識解析装置１０にて、既存の構造化データとそれに関連する文書情報を取得し、知識体系辞書作成部１２にて構造化データの同一の末端クラスに属する文書に対してキーワード抽出処理をし、重要語と重みからなる特徴ベクトルを生成し、それを末端クラスの特徴として上記取得した情報とともに分類体系知識辞書１３に格納する。未分類の文書を取得した場合、自動分類装置２０の情報抽象化部２２によりキーワード抽出処理して特徴ベクトルを求め、それを情報分類部２３にてすでに登録された各末端クラスの特徴ベクトルとマッチングを採り、一致度の高い末端クラスに割り付ける。



1

【特許請求の範囲】

【請求項１】　複数の文書を体系的に分類・整理した構造化データと、その構造化データにより分類分けされた前記複数の文書を取得し、

前記取得した各文書の特徴量を抽出するとともに、同一の末端クラスに属する文書の前記特徴量に基づいてその末端クラスの内容を特定する特徴量を決定することにより、前記構造化データの体系を解析し、

前記決定した末端クラスの特徴量と、前記構造化データ並びに前記複数の文書を関連づけて記憶手段に格納するようにした体系化知識解析方法。

【請求項２】　前記構造化データを構成する前記末端クラスと、その末端クラスまでにいたる複数の分岐点となるメタクラスの接続関係を検索し、

各クラスに対しそれと関連する上位クラス・下位クラス並びに同位クラスへのポインタを関連づけて前記記憶手段に格納することにより、前記構造化データの体系の解析をするようにした請求項１に記載の体系化知識解析方法。

【請求項３】　請求項１または２の方法を実行して得られた体系化知識を用いて未分類の文書を適当な末端クラスに関連づける分類方法であって、

処理対象の文書に対して、請求項１と同様の特徴量抽出処理を行い、その処理対象の文書の特徴量を求め、

次いで、その求めた特徴量と、請求項１により得られた各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記処理対象の文書を関連づけて前記記憶手段に格納するようにした分類方法。

【請求項４】　請求項１または２の方法を実行して得られた体系化知識を用いて未分類の文書を適当な末端クラスに関連づける分類方法であって、

複数の文書に対してそれぞれ請求項１と同様の特徴量抽出処理を行い、各文書ごとに特徴量を求め、

各文書の特徴量をクラスタリングして、特徴量の近い文書同士を一つのグループにまとめるとともに、そのグループの代表特徴量を生成し、

次いで、その求めた代表特徴量と、請求項１により得られた各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記グループを関連づけて前記記憶手段に格納するようにした分類方法。

【請求項５】　請求項３または４の方法を実行して未分類の文書を前記末端クラスに関連づけた後、

所定のタイミングでその未分類の文書の特徴量と、その関連づけられた末端クラスの特徴量に基づいて、新たな前記末端クラスの特徴量を生成するとともに、前記記憶手段の記憶内容を更新するようにした請求項１または２に記載の体系化知識解析方法。

【請求項６】　前記特徴量は、文書中に存在する重要語と、その重要語についての重みである請求項１，２，５のいずれか１項に記載の体系化知識解析方法。

2

【請求項７】　複数の文書を体系的に分類・整理した構造化データと、その構造化データにより分類分けされた前記複数の文書を取得する知識獲得手段と、

その知識獲得手段の後段に設けられ、前記取得した各文書の特徴量を抽出するとともに、同一の末端クラスに属する文書の前記特徴量に基づいて末端クラスの内容を特定する特徴量を求める知識体系辞書作成手段と、

その知識体系辞書作成手段で生成された前記末端クラスの特徴量と、前記知識獲得手段で獲得した前記構造化データ並びに前記複数の文書を関連づけて格納する記憶手段とを備えた体系化知識解析装置。

【請求項８】　請求項７に記載の体系化知識解析装置で解析して得られた体系化知識を用いて未分類の文書を適当な末端クラスに関連づける分類装置であって、

処理対象の文書を取得する情報収集手段と、

その情報収集手段で取得した所定の文書に対し特徴量抽出処理を行い、その処理対象の文書の特徴量を求める情報抽象化手段と、

その情報抽象化手段で求めた特徴量と、前記体系化知識解析装置に格納された各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記処理対象の文書を関連づけて記憶させる情報分類手段とを備えた分類装置。

【請求項９】　前記情報抽象化手段は、複数の文書を一括して処理する際に、各文書の特徴量からクラスタリングを行い、特徴量の近い文書同士を一つのグループにまとめるとともに、そのグループの代表特徴量を生成する機能を有し、

前記情報分類手段は、その代表特徴量と、各末端クラスの特徴量とのマッチングをとるものである請求項８に記載の分類装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、体系化知識解析方法及び装置並びに分類方法及び装置に関するものである。

【０００２】

【従来の技術】現在、情報入手の一手段としてインターネットが用いられている。よく知られているように、インターネットを上手に使うことにより、莫大な量と質の情報を入手することができる。そして、そのように大量に入手した情報の中から必要な情報を効率よく抽出することが必要不可欠である。また、インターネットは、世界中に存在する情報を入手することが可能になる一方、そのように大量に存在する情報の中から必要な情報を抽出するのも重要で多大な労力が必要となる。

【０００３】さらに、同様のことは、自分で各種のデータベースを作成する場合にも言える。つまり、インターネットを介して、及びまたは別の手段を介して各種の情報を取得することは比較的容易にできる。従って、何ら

かのデータベースを作成するに際し、登録する情報は集まるものの、その登録した情報をその後に検索する場合の効率を考えると、内容に応じた分類分けをする必要がある。そして、そのような分類分けをうまく行えるか否かが、その後のデータベースの使い勝手の良し悪しに顕著に反映される。

【０００４】そして、ある情報を抽出するための検索システムとしては、一般にキーワード検索が行われている。これは、入力されたキーワードをテキストデータ中に含む情報を抽出することを基本としている。しかし、単純なキーワード検索では、たまたま文書中にキーワードと同一の言語を含んでいても抽出されてしまい、検索効率が悪い。

【０００５】そこで、関連する分野を絞り込むようにしてある程度階層付けを行い、メタクラスで分岐させるツリー状の構造化データを作成し、そのツリーの最終端である末端クラスに、該当する情報を関連づけることが行われている。そして、検索しようとした場合には、そのツリーに従って、順次下位の階層に進んでいき、最終的に必用な情報を抽出するようにしたものもある。

【０００６】

【発明が解決しようとする課題】しかしながら、上記した従来の階層付け（ツリー）を行ったシステムの場合には、使用者はツリーがどのように分岐され、最終的にどのような末端クラスがあるかを予め知っている必要があるので、係るツリー構造に対する知識が十分でないと、検索効率が悪く、所望の情報を抽出することができなくなるおそれがある。また、新たに入手した情報を、すでにあるツリーの所望の末端クラスに関連づけようとした場合に、検出対象が属する分野（産業分野）である対象領域についての体系的な知識（ツリー構造）に対する知識が十分でないと、どの末端クラスに関連づければよいかわからず、間違って関連づけるおそれもあり、そうすると、その後の検索効率はさらに悪くなる。

【０００７】本発明は、上記した背景に鑑みてなされたもので、その目的とするところは、上記した問題を解決し、構造化データの状態について十分に知らなくても、未分類の書類をその内容に合った所定の末端クラスを抽出し、関連づけることができ、また、必要な情報について記載された書類を容易に検索することのできる体系化知識解析方法及び装置並びに分類方法及び装置を提供することにある。

【０００８】

【課題を解決するための手段】上記した目的を達成するために、本発明に係る体系化知識解析方法では、複数の文書を体系的に分類・整理した構造化データと、その構造化データにより分類分けされた前記複数の文書を取得し（オンライン或いはオフラインのいずれでも良い）、前記取得した各文書の特徴量を抽出するとともに、同一の末端クラスに属する文書の前記特徴量に基づいてその末端クラスの内容を特定する特徴量を決定することにより、前記構造化データの体系を解析し、前記決定した末端クラスの特徴量と、前記構造化データ並びに前記複数の文書を関連づけて記憶手段に格納するようにした（請求項１）。

【０００９】また、前記構造化データを構成する前記末端クラスと、その末端クラスまでにいたる複数の分岐点となるメタクラスの接続関係を検索し、各クラスに対しそれと関連する上位クラス・下位クラス並びに同位クラスへのポインタを関連づけて前記記憶手段に格納することにより、前記構造化データの体系の解析をするようにしてもよい（請求項２）。

【００１０】また、本発明に係る分類方法は、請求項１または２の方法を実行して得られた体系化知識を用いて未分類の文書を適当な末端クラスに関連づける分類方法であって、処理対象の文書に対して、請求項１と同様の特徴量抽出処理を行い、その処理対象の文書の特徴量を求め、次いで、その求めた特徴量と、請求項１により得られた各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記処理対象の文書を関連づけて前記記憶手段に格納するようにした（請求項３）。また、マッチングを採るに際し、その前処理として処理対象の文書が複数存在する場合には、各文書に対する特徴量を求めた後、各文書の特徴量をクラスタリングして、特徴量の近い文書同士を一つのグループにまとめるとともに、そのグループの代表特徴量を生成し、その求めた代表特徴量と、請求項１により得られた各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記グループを関連づけて前記記憶手段に格納するようにしてもよい（請求項４）。

【００１１】そして、上記した　請求項３または４の分類方法を実行して未分類の文書を前記末端クラスに関連づけた後、所定のタイミングでその未分類の文書の特徴量と、その関連づけられた末端クラスの特徴量に基づいて、新たな前記末端クラスの特徴量を生成するとともに、前記記憶手段の記憶内容を更新するようにしてもよい（請求項５）。

【００１２】そして、前記特徴量は、例えば文書中に存在する重要語と、その重要語についての重みとすることができる（請求項６）。係る場合、各請求項で記載した特徴量抽出処理は、文書中の語句からキーワードを自動的に抽出する各種のキーワード抽出システム・アルゴリズムを用いることができる。

【００１３】そして、上記した各方法を実施するために適した装置としては、例えば、複数の文書を体系的に分類・整理した構造化データと、その構造化データにより分類分けされた前記複数の文書を取得する知識獲得手段と、その知識獲得手段の後段に設けられ、前記取得した各文書の特徴量を抽出するとともに、同一の末端クラスに属する文書の前記特徴量に基づいて末端クラスの内容

を特定する特徴量を求める知識体系辞書作成手段と、その知識体系辞書作成手段で生成された前記末端クラスの特徴量と、前記知識獲得手段で獲得した前記構造化データ並びに前記複数の文書を関連づけて格納する記憶手段とを備えた体系化知識解析装置（請求項７）とすることができる。

【００１４】そして、係る請求項７に記載の体系化知識解析装置で解析して得られた体系化知識を用いて未分類の文書を適当な末端クラスに関連づける分類装置では、処理対象の文書を取得する情報収集手段と、その情報収集手段で取得した所定の文書に対し特徴量抽出処理を行い、その処理対象の文書の特徴量を求める情報抽象化手段と、その情報抽象化手段で求めた特徴量と、前記体系化知識解析装置に格納された各末端クラスの特徴量とのマッチングをとり、一致度の高い末端クラスに前記処理対象の文書を関連づけて記憶させる情報分類手段とを備えるように構成することができる（請求項８）。そして、前記情報抽象化手段は、複数の文書を一括して処理する際に、各文書の特徴量からクラスタリングを行い、特徴量の近い文書同士を一つのグループにまとめるとともに、そのグループの代表特徴量を生成する機能を有し、前記情報分類手段は、その代表特徴量と、各末端クラスの特徴量とのマッチングをとるものとしてもよい（請求項９）。

【００１５】＊用語の定義

「構造化データ」は、例えばツリー構造（階層構造）等のように特定の分野における文書情報を体系的に分類・整理する際に用いる分類の仕様・体系そのものを示す情報である。また、「体系化知識」は、構造化データがどのような意味・視点等にたって分類されているかを示す知識であり、本発明では、少なくとも上記構造化データを構成する末端クラス（それより下位に分岐されるクラスがなく、文書データが接続されている）がどのような内容の文書を接続すれば良いかを示す知識であればよい。

【００１６】

【発明の実施の形態】まず、本実施の形態が取得する構造化データの一例を示すと、図１のようになっている。図示の例では、各種のニュースをその内容に応じて分類分けしている。すなわち、最上位のメタクラスである「Ｎｅｗｓ」の下に「Ｓｐｏｒｔｓ」，「Ｅｃｏｎｏｍｉｃｓ」……等の各分野に分けられ、さらに各分野はその内容に応じて細分類化される。そして、分類分けができないものが末端クラス（図中ハッチングで示す）となり、各末端クラスに該当する書類（ドキュメント）が関連づけられている。なお、当然のことながら各末端クラスに関連づけられた各種の書類は、そのメタクラスの上位に接続されたすべてのメタクラスの要件・内容を満たすものである。

【００１７】そして、図示するようなツリー構造で各対象領域についてその分類構造を体系的に表わしたものは、例えばインターネット上の検索ソフト（サイト）に予め構築されていることが多く、ある情報を検索したい場合には、通常係るサイトにアクセスし、関連づけられた各メタクラスを順番に辿りながら、目的とする末端クラスに到達し、そこに属する書類を閲覧したり一括してダウンロードしたり、他のサイトに飛んだりすることになる。また、そのようなツリー構造を作成した人はもちろんのこと、他の人も新たな書類を該当する末端クラスに関連づけることもある。

【００１８】そこで本形態では、上記したすでに存在する構造化データ及びそれに関連づけられた書類を取得し、その構造化データの解析を行うとともに、その解析により取得した分類体系化知識に基づいて新たに入手した情報をその構造化データの所定の末端クラスに関連づけを行うことができるようにしている。そして、係る処理を行うための具体的なシステム構成の一例を示すと、図２のようになっている。

【００１９】同図に示すように、本システムは、体系化知識解析装置１０と、その体系化知識解析装置１０にて解析した結果に基づいて、新たに入手した情報（書類）を分類整理する自動分類装置２０と、上記した各装置１０，２０により構築したデータベースに対して情報検索を行う検索装置３０の各実施の形態を備えている。

【００２０】まず、体系化知識解析装置１０は、入力側に分類体系知識獲得部１１を備え、インターネットなどを介して既存の構造化データを取得する。この時、取得するのは図１に示すようなツリー構造の構造化データ自体と、それに関連づけられた書類である。

【００２１】そして、そのようにして取得したデータを次段の知識体系辞書作成部１２に与える。この知識体系辞書作成部１２は、解析対象の末端クラスに関連づけられた書類（文書）、すなわち、実例データを受け取り（ＳＴ１）、各書類ごとに特徴ベクトルを生成する（ＳＴ２）。

【００２２】この書類の特徴ベクトルの生成処理の一例を示すと、まず処理対象の文書中の重要語リストを抽出する。この重要語リストの抽出は、例えば文書中に存在するすべての名詞を抽出し、その名詞の出現回数などに基づいて各名詞に重み付けを行い、重みの大きいもののうち上から所定数を重要語リストとして抽出する等、データベース作成時に用いられる各種のキーワード自動抽出アルゴリズムを用いることができる。そして、その抽出された重要語と重みを関連づけたデータをその文書についての特徴ベクトルとする。係る特徴ベクトルは、その末端クラスに属するすべての書類について行う。従って、末端クラスに関連づけられた種類の数だけ文書の特徴ベクトルが生成される。

【００２３】次いで、それら生成されたすべての特徴ベクトルの平均化処理を行い、その末端クラスに属するす

べての書類の平均特徴ベクトルを求め、それをその末端クラスの特徴ベクトルとする（ＳＴ３，ＳＴ４）。上記した平均化処理は、例えばステップ２で求めた同一の末端クラスに属するすべての文書の特徴ベクトルは、重要語とその重みにより構成されているので、各文書で抽出された重要語をすべて拾い出すとともに、それについて付された重みの平均値をとる。単純な平均値の求め方としては、同一の重要語の重みをすべて加算し、その加算値を末端クラスに属する文書数で割ることにより各重要語の末端クラスにおける重みが求められる。

【００２４】そして、少なくとも１つの文書から抽出された重要語はすべてその末端クラスにおける重要語として特徴ベクトルに反映させるようにしてもよいし、そのように平均化処理をして得られた重みが一定の値以上の重要語を最終的に残してもよいし、或いは、重みの大きい語句から所定数を最終的に重要語（キーワード）として残すようにしてもよく、各種の方式をとることができる。さらに、各文書で同一の重要語が抽出された場合には、係る重要語はその末端クラスに属する書類を特徴づけるものとしてより重要であるといえ、１または少数の書類にのみ抽出された重要語は、その末端クラスに属する書類を特徴づけるものとしてはさほど重要でないといえる。従って、各書類における重みとともに、同一の末端クラスに属する書類のなかで、抽出された数の多い重要語の重みを重くするように処理してもよい。

【００２５】そして、そのようにして得られた末端クラスの特徴ベクトルは、例えば図４に示すようなデータ構造となる。ここで、グループＮｏは、各末端クラスを特定する番号であり、ＫＷ数はそこで抽出された重要語の数であり、ＫＷｊは、具体的に抽出された重要語であり、ｗｊは、ＫＷｊについての重みである。そして、上記した処理を取得したすべての末端クラスについて求める。そして、そのようにして求めた各末端クラスについての特徴ベクトルを、次段の分類体系知識辞書１３に格納する。

【００２６】また、知識体系辞書作成部１２は、上記した各末端クラスの特徴ベクトルを生成する機能に加え、図１に示すようなツリー構造をデータ化する機能も有している。すなわち、図１に示すようなツリー構造は、図５に示すように、最上位（図示の例では「Ｎｅｗｓ」）のレベル０から順に下位にいくに従ってレベルが１ずつ増えていくとする。そして、各メタクラス・末端クラスについてレベル付けを行うとともに、クラス間の接続関係を求める。そして、各クラスの接続先（ポインタ）を見つけるとともに、両者のレベルの大小関係を比較し、接続先が上位／下位／同位かを判断する。さらに、そのクラスがメタクラスか末端クラスかの弁別も行う。

【００２７】そして、係る処理を行った結果、図６に示すような各クラスについての「クラス名・レベル・接続先を示すポインタ及び末端クラスか否かのフラグ」を関連づけたテーブルを作成する。さらに、末端クラスの場合には、それより下位のクラスがないため、その下位ポインタの欄には、その末端クラスについて求めた特徴ベクトルを格納したアドレスを下位のポインタとして登録している。そして、そのようにして形成したテーブルを、分類体系知識辞書１３に格納するようにしている。

【００２８】さらに本形態では、体系化知識解析装置１０には、知識更新部１４を備え、所定のタイミングで分類体系知識辞書１３に格納した末端クラスの特徴ベクトルを更新するようにしている。具体的には、後述する自動分類装置２０により新たに分類整理して追加された書類が所定数たまった場合に、それら追加された書類を含めてその時存在する末端クラスに属する書類に対して、上記したのと同様の処理を実行し新たな特徴ベクトルを生成し、書き換える。

【００２９】すなわち、既存の末端クラスに付されている特徴ベクトルを構成する重要語（重み付き）と、新たに入手した情報から得られた特徴ベクトル（後述するトピック情報）を構成する重要語（重み付き）の和集合を求め、その和集合を該当する末端クラスの新たな特徴ベクトルとする。そして、各重要語の重みは、既存の特徴ベクトルの重要語の重みと新たに入手した情報の重要語の重みの加重平均により求めるようにしている。なお、上記した和集合を構成する重要語が、元の特徴ベクトルにない場合には、そのない方の特徴ベクトルにおける当該重要語の重みは０として加重平均を求めることにしている。

【００３０】一方、自動分類装置２０は、図７に示すような処理フローを実行する機能を備えており、具体的には、入力側に情報収集部２１を有し、その情報収集部２１は、インターネットなどを介して未整理の文書情報（書類）を取得し、次段の情報抽象化部２２に与える。この時入手する書類としては、単一でもよいし複数でもよい（ＳＴ１１）。

【００３１】この情報抽象化部２２は、図７におけるステップ１２，１３を実行するもので、まず、取得したすべての書類に対し、書類ごとの特徴ベクトルを生成する（ＳＴ１２）。係る生成処理は、知識体系辞書作成部１２における処理と同様のものを用いることができる。次いで、複数の書類を取得した場合には、各書類の特徴ベクトル（重要語とその重み情報）についてクラスタリングを行い、類似する物同士をグループ化する。次いで、各グループを代表する特徴ベクトル（代表特徴ベクトル）を求める。この代表特徴ベクトルは、例えば知識体系辞書作成部１２において末端クラスの特徴ベクトルを生成したのと同様に、そのグループに属する書類についての特徴ベクトルの平均値を求めることにより簡単に生成できる。もちろん、他の手法により求めてもよい。ここまでの処理がステップ１３であり、この処理を実行して得られたグループを構成する書類と、その代表特徴ベ

クトル情報を次段の情報分類部２３に送る。

【００３２】情報分類部２３では、分類体系知識辞書１３に格納された各末端クラスの特徴ベクトルを読み出すとともに、与えられた各グループについての代表特徴ベクトルを比較し、マッチングをとる（ＳＴ１４）。この時、比較する両特徴ベクトルを構成するキーワード数を同じにすべく、重みの大きい重要語からｋ個を抽出してグループについてのトピック情報を求め、その重要語と重みに基づいてマッチングをとり、最も一致する特徴ベクトルの末端クラスにそのグループを構成する未知の書類を割り付けることを決定する（ＳＴ１５）。

【００３３】ここで、トピック情報のデータ構造としては、例えば図８に示すようになっており、図４に示す各末端クラスの特徴ベクトルと同様のデータ構造で、違うのは、図４のものが先頭がクラス名であるのに対し、図８のものは未連結なためそのグループ番号が先頭である点である。そして、マッチング処理により、特徴ベクトルが最も近い末端クラスが決定されると、図９に示すようなテーブルのうち、グループ番号，グループを構成する文書・書類が格納された先頭のデータレコードへのポインタ並びにステップ１５で決定された関連づけられる末端クラス名を登録する。

【００３４】また、階層レベルや、その末端クラス名が接続される上位クラスや同位クラスへのポインタは、関連づけられた分類クラス名が決まると一義的に決まるので、分類体系知識辞書１３にアクセスして係る階層レベルや各所へのポインタデータを抽出し、登録する。なお、特徴ベクトルのマッチング処理は公知の各種のものを用いることができるので、その詳細な説明を省略する。

【００３５】そして、上記のように未知のグループの割付（関連先の末端クラスの決定）が終了したならば、その結果を出力表示すべくデータを加工する。それがステップ１６である。つまり、グループが複数存在する場合には、相関がとれずにバラバラになっており、しかも、本形態では、未知情報を入手する都度、構造化データにおける末端クラスの特徴ベクトルを更新するのではないので、次の更新処理をするまでに、何回か上記した入手した未知の書類に対するグループ化に基づくトピック情報（特徴ベクトル）の生成に伴う分類処理を行っている場合には、同一の末端クラスに属するグループが複数存在することもある。従って、それらを統計だてて出力表示するために、データを加工するようにしている。

【００３６】そして、そのステップ１６の具体的な処理は、図１０に示すようになる。すなわち、ステップ１５を実行して処理対象のすべてのグループの割り付けが終わったならば、上記した図８，図９に示す各グループの割付結果のデータを取得し、それを図１１（Ａ）に示すような出力データ（Ａ）に変換する（ＳＴ１６ａ）。つまり、同一グループについての図８，図９に示すデータのうち、出力データ（Ａ）の各欄に該当するものを登録することにより行う。この時、同一の末端クラスに属するグループが複数ある場合には、その末端クラスについての出力データ（Ａ）に登録する。これにより、同一の末端クラスに属する書類は、１つのデータレコードにまとめられる。

【００３７】次に、出力データ（Ａ）の集合に対し、同位クラスへのポインタの項目に基づいてソートし、それより各上位クラスを抽出する。この上位クラスの抽出は、例えば上位クラスへのポインタに基づいて容易に行える。この抽出に従い、図１１（Ｂ）に示すような出力データ（Ｂ）を生成する（ＳＴ１６ｂ，１６ｃ）。上位クラスは当然のことながらそれに続く下位クラスが存在する（上位クラスの抽出のもとになったもの）。

【００３８】これにより、末端クラスから一つ上の階層レベルに属する上位クラスについてのデータが生成される。そして、その上位クラスもさらにその上位クラスが存在することがあるので、生成された各出力データ（Ｂ）で、共通の上位クラスがあるか否かを判断し（ＳＴ１６ｄ）、ある場合には、ステップ１６ｃに戻りさらにその上位クラスについての出力データ（Ｂ）を生成する。

【００３９】そして、係る分類クラスについての出力データ（Ａ），（Ｂ）を次段の結果表示部２４に与える。結果表示部２４では、取得した出力データに対し、階層レベルをキーにソートし、図１２に示すような出力用のデータ構造からなるデータを作成し、それに基づいて、出力装置２５に結果を表示する。具体的には、例えば図１３に示すように、構造化データ（ツリー構造）とともに、各末端クラスに属するトピック情報を表示したり、図１４に示すように具体的な重要語（キーワード）を表示したりすることができる。

【００４０】なお、上記した処理をしてもトピック情報と既存の末端クラスの特徴ベクトルとの一致度が低く、どれとも関連づけられない場合もある。係る場合には、例えば上記の表示された構造化データをみながら、マニュアル操作により、妥当なメタクラスの下に末端クラス名を作成し、それを新しい知識として既存の知識体系に追加するようにしている。

【００４１】また、情報分類部２３は、分類体系知識辞書１３に格納された構造化データとそれに関連づけられる書類及び体系化知識（末端クラスの特徴ベクトル）や、新たに入出した書類（グループ）等の情報を全文検索データベース３１に格納するようにしている。

【００４２】検索装置３０は、上記全文検索データベース３１と、情報検索部３２を備えており、キーボードなどの入力装置３３を介して与えられた検索キーに基づいて情報検索部３２が全文検索データベース３１をアクセスし、該当する文書を抽出するようにしている。そして、その抽出結果は、結果表示部２４を介して表示装置

11

２５に表示するようにしている。

【００４３】そして、この情報検索部３２における検索処理としては、従来の全文一括のキーワード検索と同様に、全文検索データベース３１中に登録された各書類のテキストデータをすべてサーチし、文書中に入力されたキーワードを含む文書を抽出することができる。また、上記した特徴ベクトルを利用して、必要な情報を有する末端クラスを抽出し、それに属する書類を表示したり、一括してダウンロードしたり、目次などを表示して所定の書類を選択することなどができるようになっている。そして、具体的な検索方式としては、例えば特徴ベクトルとして、検索したい情報について含まれると予想する重要語とその重みを関連づけたものを複数個入力する。そして、入力した重要語と重みと、すでに登録された各末端クラスについての特徴ベクトルとのマッチングを採り、最も近いものを該当する末端クラスと決定し、抽出することができる。

【００４４】

【発明の効果】以上のように、本発明に係る体系化知識解析方法及び装置並びに分類方法及び装置では、端末クラスに関連づけられた文書の特徴量を抽出し、同一の端末クラスに損する文書の特徴量からその端末クラスの特徴量を決定するため、具体的な構造化データの状態について十分に知らなくても、未分類の書類をその内容にあった所定の末端クラスを抽出し、関連づけることができる。

【００４５】また、そのように取得した構造化データと文書に、解析した特徴量を関連づけて登録するため、その後に必要な情報を検索する場合には、特に構造化データの状態を知らなくても、係る特徴量を検索キーにしてサーチすることにより、必要な書類が関連づけられている末端クラスを抽出できる。つまり、単純なキーワード検索よりも高精度で、不要な情報を抽出する可能性を低く抑えることができる。

【図面の簡単な説明】

【図１】構造化データの一例を示す図である。

【図２】本発明の好適な一実施の形態を示すブロック図である。

12

【図３】知識体系辞書作成部の機能の一部を示すフローチャートである。

【図４】特徴ベクトルのデータ構造を示す図である。

【図５】構造化データを解析する際のポインタを説明する図である。

【図６】知識体系辞書作成部で解析して得られた各クラスの接続関係を登録する際のデータ構造を示す図である。

【図７】分類装置の機能を説明するフローチャートである。

【図８】クラスタリングにより得られた結果を格納する際のデータ構造の一例を示す図である。

【図９】分類クラスの割り付けにより得られた結果を格納する際のデータ構造の一例を示す図である。

【図１０】上位クラスの決定アルゴリズムを説明するフローチャートである。

【図１１】出力用のデータレコード仕様を示すデータ構造図である。

【図１２】出力用のデータ構造を示す図である。

【図１３】出力表示例を示す図である。

【図１４】出力表示例を示す図である。

【符号の説明】

１０　体系化知識解析装置
１１　分類体系知識獲得部
１２　知識体系辞書作成部
１３　分類体系知識辞書
１４　知識更新部
２０　自動分類装置
２１　情報収集部
２２　情報抽象化部
２３　情報分類部
２４　結果表示部
２５　表示装置
３０　検索装置
３１　全文検索データベース
３２　情報検索部
３３　入力装置

【図４】

クラス名	KW数	KW1:w1	KW2:w2	…	KWm:wm
Football	60	nfl:0.9	wlaf:0.9	…	goods:0.5

KWj：j番目のキーワード（重要語）
wj　：j番目のキーワードの重み（重要度）

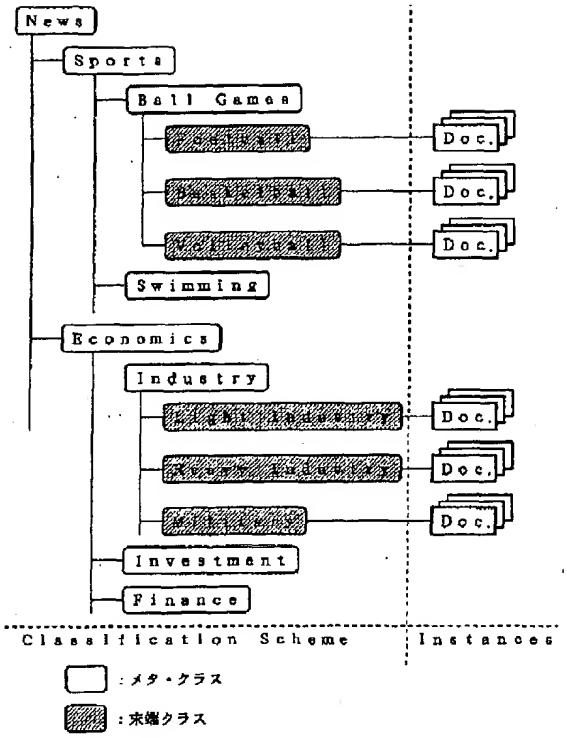
末端クラス（j）の特徴ベクトルの構成とその例

【図８】

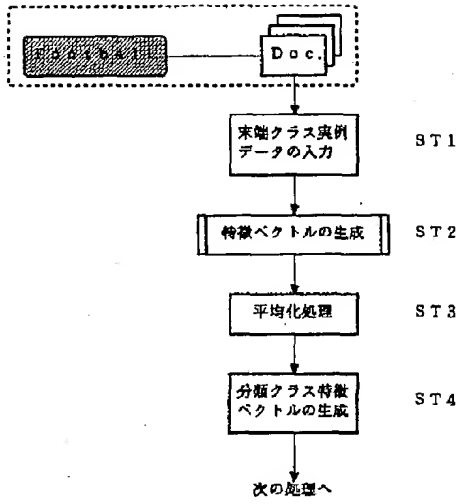
グループNo.	KW数	KW1:w1	KW2:w2	…	KWk:wk

グループのトピック情報のデータレコード仕様

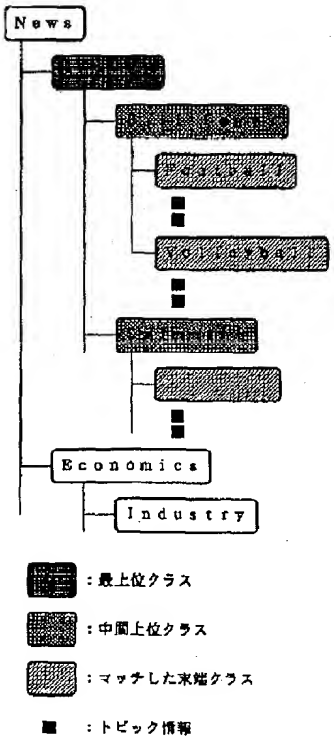
【図１】



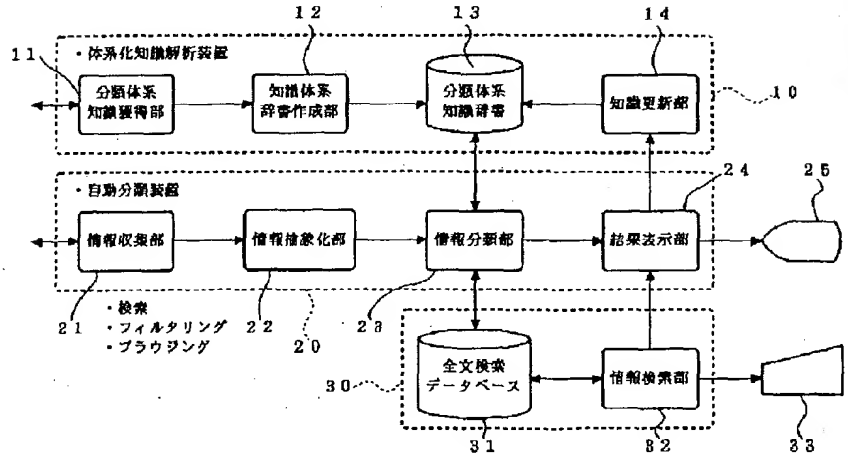
【図３】



【図１３】



【図２】



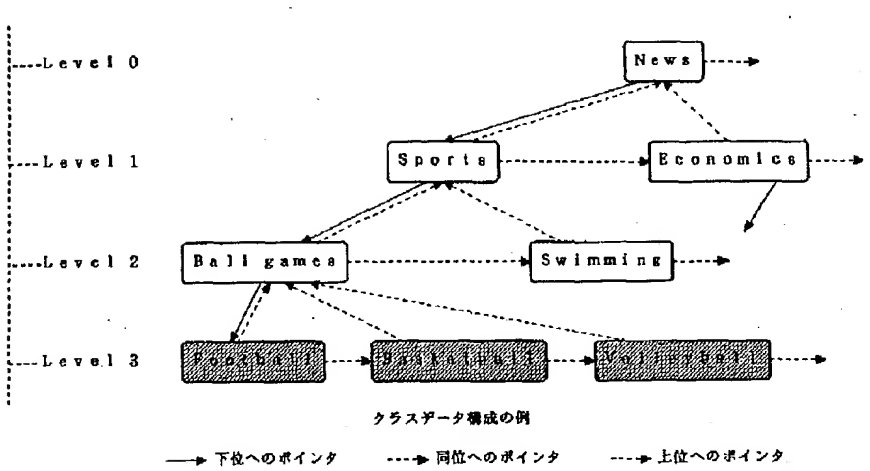
【図９】

グループNo.	Pt	末端分類クラス名	階層レベル	上位クラスへのポインタ	同位クラスへのポインタ

Pt：グループトピックデータレコードへのポインタ

分類クラスへの割り付けの結果に関するデータレコード仕様

【図５】



クラスデータ構成の例

【図６】

クラス名	階層レベル	上位クラスへのポインタ	下位クラスへのポインタ	同位クラスへのポインタ	Flag
Sports	1	P_news	P_ballgames	P_economics	0
Ball game	2	P_sports	P_football	P_swimming	0
Football	3	P_ballgames	P_feature	P_basketball	1
…	…	…	…	…	…

Flag　　：非末端クラス＝０；　末端クラス＝１
P_feature：該当特徴ベクトルへのポインタ

クラス（ｉ）の属性構成とその例

【図１１】

(A)

末端分類クラス名	階層レベル	上位クラスへのポインタ	同位クラスへのポインタ	グループ数（m）	Pt1	…	Ptm

出力用のデータレコード仕様（A）

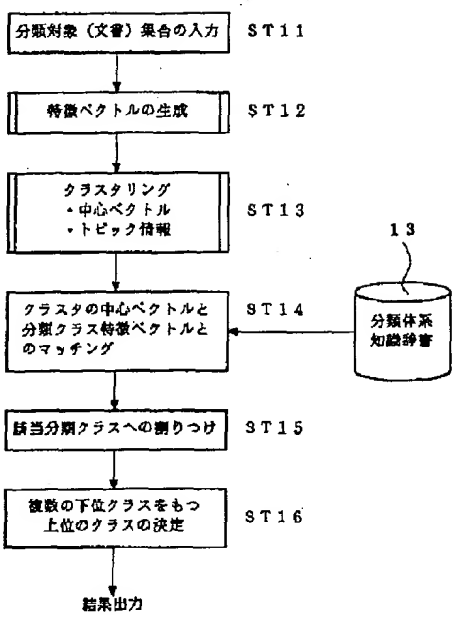
(B)

分類クラス名前	階層レベル	上位クラスへのポインタ	下位クラスへのポインタ	同位クラスへのポインタ	Flag

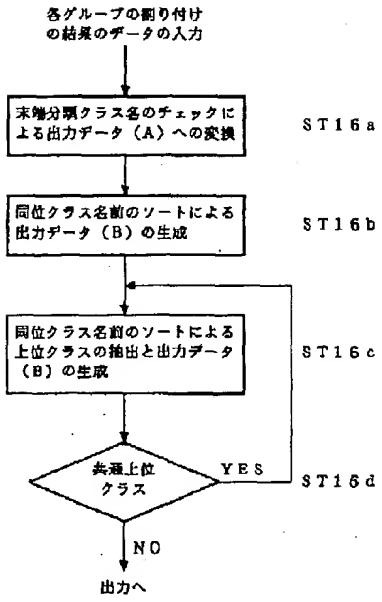
Flag　：非末端クラス＝０；　末端クラス＝１

出力用のデータレコード仕様（B）

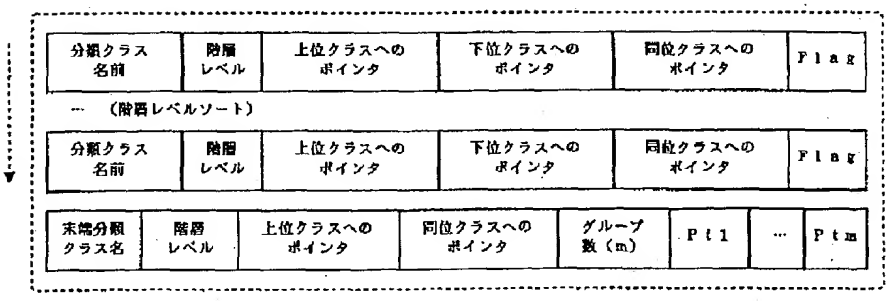
【図７】



【図１０】

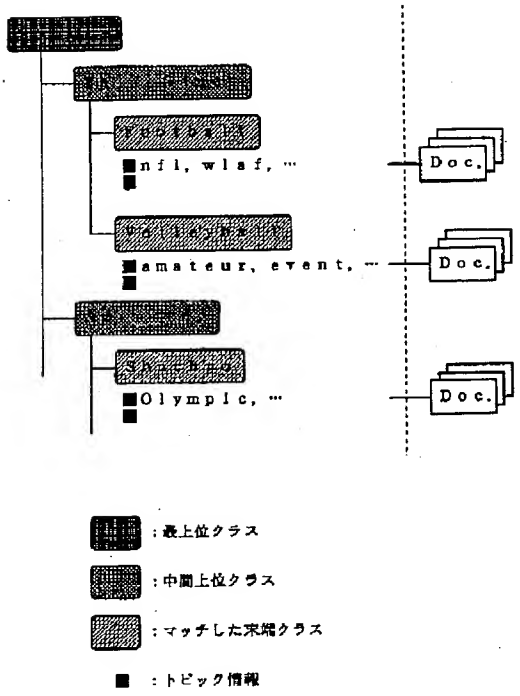


【図１２】



出力用のデータ構造

【図１４】



フロントページの続き

(72)発明者　十河　太治
京都府京都市右京区花園土堂町10番地　オムロン株式会社内

(72)発明者　澤田　晃
京都府京都市右京区花園土堂町10番地　オムロン株式会社内